Fine-Scale Haplotype Structure Reveals Strong Signatures of Positive Selection in a Recombining Bacterial Pathogen

Arnold, Brian

2020-02

Arnold , B , Sohail , M , Wadsworth , C , Corander , J , Hanage , W P , Sunyaev , S & Grad , Y H 2020 , ' Fine-Scale Haplotype Structure Reveals Strong Signatures of Positive Selection in a Recombining Bacterial Pathogen ' , Molecular Biology and Evolution , vol. 37 , no. 2 , pp. 417-428 . https://doi.org/10.1093/molbev/msz225

http://hdl.handle.net/10138/317197
https://doi.org/10.1093/molbev/msz225

Fine-Scale Haplotype Structure Reveals Strong Signatures of Positive Selection in a Recombining Bacterial Pathogen

Brian Arnold,*[,1,2] Mashaal Sohail,[†,3,4] Crista Wadsworth,[‡,5] Jukka Corander,[6,7] William P. Hanage,[2] Shamil Sunyaev,[3,4] and Yonatan H. Grad[5,8]

[1]Division of Informatics, Faculty of Arts and Sciences, Harvard University, Cambridge, MA

[2]Center for Communicable Disease Dynamics, Harvard T. H. Chan School of Public Health, Boston, MA

[3]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA

[5]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA

[6]Department of Biostatistics, University of Oslo, Oslo, Norway

[7]Department of Computer Science, Helsinki Institute for Information Technology HIIT, University of Helsinki, Helsinki, Finland

[8]Division of Infectious Diseases, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

[†]Present address: National Laboratory of Genomics for Biodiversity (UGA-LANGEBIO), CINVESTAV, Irapuato, Guanajuato, Mexico

[‡]Present address: Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY

***Corresponding author:** E-mail: barnold@g.harvard.edu.

Associate editor: Sergei Kosakovsky Pond

Abstract

Identifying genetic variation in bacteria that has been shaped by ecological differences remains an important challenge. For recombining bacteria, the sign and strength of linkage provide a unique lens into ongoing selection. We show that derived alleles <300 bp apart in *Neisseria gonorrhoeae* exhibit more coupling linkage than repulsion linkage, a pattern that cannot be explained by limited recombination or neutrality as these couplings are significantly stronger for non-synonymous alleles than synonymous alleles. This general pattern is driven by a small fraction of highly diverse genes, many of which exhibit evidence of interspecies horizontal gene transfer and an excess of intermediate frequency alleles. Extensive simulations show that two distinct forms of positive selection can create these patterns of genetic variation: directional selection on horizontally transferred alleles or balancing selection that maintains distinct haplotypes in the presence of recombination. Our results establish a framework for identifying patterns of selection in fine-scale haplotype structure that indicate specific ecological processes in species that recombine with distantly related lineages or possess coexisting adaptive haplotypes.

Key words: **evolutionary genetics, microbiology, computational biology, simulation, adaptation, linkage.**

Introduction

Bacteria colonize diverse environments in which they experience challenges that leave distinct patterns in their genomes (Shapiro et al. 2009). As these environmental pressures are often unknown, genomic signatures of selection can help guide our understanding of bacterial ecology and evolution.

Unexpected patterns of linkage remain an important indicator of recent positive or negative selection, as many bacteria undergo frequent recombination (Vos and Didelot 2009) that unlinks loci. *Neisseria gonorrhoeae*—a clinically important pathogen and an urgent public health concern due to growing incidence and antibiotic resistance—recombines extensively with closely and distantly related species and has been previously described as "panmictic" and "freely recombining" (O'Rourke and Stevens 1993; Smith et al. 1993). However, like in many bacterial species, proximate loci in *N. gonorrhoeae* exhibit stronger linkage than more distant loci (Arnold et al. 2018). This is as expected under neutral models that show neighboring polymorphisms are transferred together on recombination tracts, which range from tens to thousands of base pairs (Arnold et al. 2018; Lin and Kussell 2019). Many bacteria, including *N. gonorrhoeae*, also frequently exchange alleles with other species, thereby introducing clusters of linked mutations.

Although neutral processes may thus elevate background levels of linkage for proximate loci, selection may also shape fine-scale haplotype structure, especially since the strength of selection acting on associations between neighboring loci may easily exceed the rate at which recombination breaks them. Consequently, methods to detect these signatures will aid understanding of microbial adaptation to ecological pressures and thus may inform the development of measures to control bacterial pathogens.

Here, we develop an approach to study the interaction between selection and recombination within short genetic distances. We compare the sign and strength of linkage

Article

between nonsynonymous (NSyn) derived alleles and synonymous (Syn) derived alleles, matched for relative recombination rate and allele frequency to control for population structure. Positive associations between derived alleles indicate coupling linkage, whereas negative associations indicate repulsion linkage. Using three large population genomic data sets of *N. gonorrhoeae*, we find that proximate loci generally exhibit an excess of coupling linkage, which cannot be explained by limited recombination. Intriguingly, NSyn alleles <300 bp apart show significantly stronger couplings than Syn alleles. We found that this pattern is driven by a subset (<10%) of exceptionally diverse genes, many of which exhibit evidence of interspecies recombination and an excess of intermediate-frequency alleles. Included among these genes are many metabolic proteins, and membrane proteins and the Mtr operon that we previously showed contain horizontally transferred alleles that confer antibiotic resistance (Wadsworth et al. 2018). Many types of selection affect patterns of linkage, but extensive simulations show that only two specific forms of positive selection can create these signatures: adaptive horizontal gene transfer (HGT) and balancing selection, defined here as spatially or temporally variable.

Directional linkage measures applied to recombining bacteria that exchange alleles with other species can thus provide a wealth of information about selection within genes, information that may go undetected by approaches that aim to detect linkage outliers between distant polymorphisms (Cui et al. 2015; Skwark et al. 2016; Schubert et al. 2019) or by conservative tests for selection that require more NSyn variation compared with Syn variation ($dN/dS > 1$).

Results

Linkage Metrics

To quantify the strength of linkage between allele pairs, we calculated the squared correlation coefficient $r^2 = D^2/p_i(1 - p_i)p_j(1 - p_j)$, where $D = p_{ij} - p_ip_j$, or the probability that derived mutations at site i and j occur together (p_{ij}) minus the random expectation based on their individual frequencies p_i and p_j (Hill and Robertson 1968).

We also considered the correlation coefficient r to measure the sign of linkage between alleles as it contains important information about haplotype structure. Values of r^2 range from 0 to 1, whereas r ranges from −1 to 1. Positive values of r indicate that alleles tend to cooccur (coupling phase), whereas negative values indicate that alleles tend to reside on different genetic backgrounds (repulsion phase). The sign of r depends on the alleles one chooses to pair at two polymorphic loci, and we quantified r between derived alleles as in Takahasi and Innan (2008) by inferring the ancestral allele using one or several outgroups (with similar results; Materials and Methods). Derived alleles represent de novo mutations in *N. gonorrhoeae* or haplotypes imported from diverged populations that harbor alleles not found in the outgroups.

In a single recombining population under neutrality, the sample mean of r ($\bar{r}$) between allele pairs separated by any distance should be near zero from equal amounts of coupling and repulsion linkage (Hill and Robertson 1968). Less recombination between close alleles increases the variance of D (and the strength of linkage as measured by r^2), but these values of D take on positive or negative values with equal frequency (Ohta and Kimura 1969, 1971) such that $\bar{r} \approx 0$ (supplementary fig. S1, Supplementary Material online).

However, a variety of evolutionary processes may create coupling linkage, or positive r, including neutral dynamics such as population structure and interspecies HGT (Martin et al. 2006), or processes involving selection such as hitchhiking, balancing selection, or positive epistasis (antagonistic epistasis between deleterious mutations, or synergistic epistasis between beneficial mutations; Eshel and Feldman 1970; Thomson 1977; Charlesworth et al. 1997). Repulsion linkage, or negative r, is caused by clonal interference between mutations or negative epistasis (synergistic epistasis between deleterious mutations, or antagonistic epistasis between beneficial mutations; Hill and Robertson 1968; Sohail et al. 2017). If any of these evolutionary processes that create coupling or repulsion linkage are weak compared with the amount of recombination, r will remain near zero as recombination rapidly breaks nonrandom associations.

Neighboring Polymorphisms in *N. gonorrhoeae* Exhibit an Excess of Coupling Linkage

Using three independent genomic data sets of clinical isolates collected from the United Kingdom (UK; $n = 214$), New Zealand (NZ; $n = 148$), and the United States (US; $n = 149$), we tested the hypothesis that *N. gonorrhoeae* is a single recombining population under neutrality using these pairwise linkage metrics. We first analyzed Syn single-nucleotide polymorphisms (SNPs) that, if neutral in effect, should provide a less biased view of recombination and population structure. Linkage between Syn alleles was generally low as expected due to extensive recombination, and r^2 approached near-zero values as SNPs became separated by distances longer than ~3 kb (fig. 1*A*), in agreement with previous work (Arnold et al. 2018). However, for close SNPs that exhibited stronger linkage, the sign of r became systematically positive (fig. 1*B*); neighboring Syn alleles were coupled more often than expected in a single neutral population, particularly for alleles separated by <300 bp (fig. 1*B*).

NSyn Couplings Stronger Than Syn Couplings between Neighboring SNPs

Assuming NSyn SNPs are the actual targets of selection and the excess of short-range Syn couplings is driven by hitchhiking, we directly examined the role of selection in shaping this pattern by comparing the degree of coupling and repulsion linkage between Syn and NSyn alleles. Specifically, we evaluated whether r between NSyn derived alleles (r_N) differed significantly from r between Syn derived alleles (r_S). Because recombination breaks linkage and brings the magnitude of associations (positive and negative) closer to zero, we confirmed that selection, not variation in recombination, drives differences between r_N and r_S. Briefly, we confirmed that any differences between r_N and r_S remained after controlling for genetic distance, allele frequencies, and local recombination

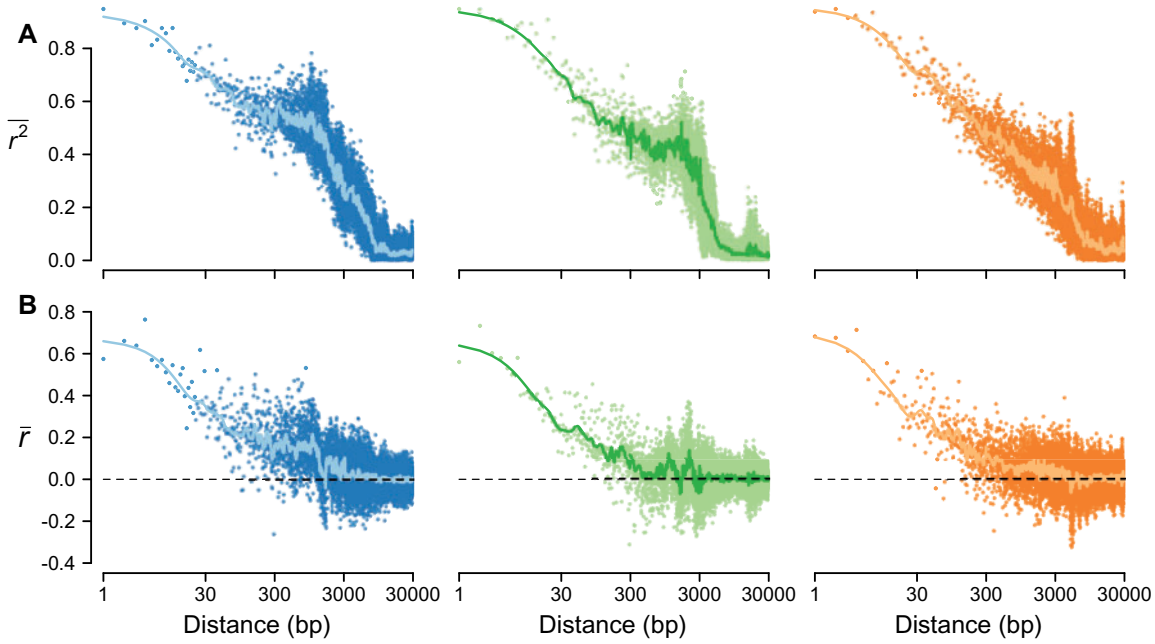

FIG. 1. Patterns of linkage in *Neisseria gonorrhoeae*. (*A*) Pairwise linkage between Syn SNPs, as measured by r^2, reached very low levels for SNPs separated by >3 kb. (*B*) However, when measured as *r*, pairwise linkage showed an excess of couplings between close SNPs that break down into random associations for distantly spaced SNPs >3 kb apart. Each point represents the mean value of r^2 or *r* for all SNP pairs separated by a number of base pairs shown in the *x*-axes, and lines are smoothing splines. Results are shown for UK (blue), NZ (green), and US (orange).

rates, all of which may cause particular alleles to have different opportunities for recombination.

As the probability that recombination unlinks two alleles varies with distance (fig. 1*A*), we first binned allele pairs by distance intervals of 100 bp. If one allele category had more observations (e.g., more Syn than NSyn SNPs), we subsampled it 100 times to estimate variability. If NSyn and Syn alleles experienced similar selective pressures, r_N would be similar to r_S within each bin, but figure 2 shows that for genomic distances <~3 kb, r_N was significantly different from r_S. Intriguingly, for distances of <~300 bp where we observed the strongest coupling linkage for Syn alleles (fig. 1*B*), r_N was greater than r_S, indicating that NSyn alleles exhibit even stronger couplings (fig. 2).

NSyn Couplings Driven by Intermediate-Frequency Alleles

To further understand the evolutionary forces shaping this excess of NSyn couplings between alleles within ~300 bp, we controlled for potential differences in recombination between NSyn and Syn alleles by considering sample frequencies, which differ between categories (supplementary fig. S2, Supplementary Material online). Rare alleles are generally younger than those at higher frequencies and have had less opportunity for recombination to break down their linkage. Although we matched allele frequencies across the entire spectrum (below), we first considered only rare alleles to study how negative selection shapes patterns of $r_N - r_S$, as rare polymorphisms are enriched for deleterious mutations. Negative selection is a pervasive evolutionary force in bacteria (Hughes 2005) that will skew NSyn alleles toward rare frequencies, make them younger than nearly neutral alleles at the same frequency, and create clonal interference (or repulsion linkage) when the strength of this selection exceeds the rate of recombination (Hill and Robertson 1968; Felsenstein 1974; McVean and Charlesworth 2000).

Focusing only on neighboring SNPs within 300 bp, when we first considered alleles <1% frequency, which comprised ~25% of all NSyn SNPs, we found that $r_N - r_S$ was significantly negative between neighboring alleles—indicating an excess of NSyn repulsion linkage (fig. 3*A*). This is expected under a model of clonal interference between deleterious mutations that experience less recombination, but negative selection with synergistic epistasis may also play a role (Eshel and Feldman 1970; Sohail et al. 2017). Nonetheless, as we observed an excess of repulsion linkage between rare NSyn alleles, negative selection is an unlikely explanation for the overall excess of coupling linkage between proximate NSyn alleles (fig. 2).

We found a significant excess of NSyn couplings, or positive $r_N - r_S$, for common alleles at 20–80% frequency that are nearly neutral or experiencing positive selection (fig. 3*A*; $P \leq 1.3 \times 10^{-7}$ by Wilcoxon rank-sum test). We again controlled for potential differences in recombination between NSyn and Syn alleles by further binning these intermediate-frequency alleles into 10% frequency intervals, as higher frequency alleles (that may consist primarily of Syn mutations) are generally older and have had more time to experience recombination. Values of $r_N - r_S$ for alleles within each interval were generally positive (supplementary fig. S3*A*, Supplementary Material online), and meta-analysis across all intervals using Stouffer's method indicated that

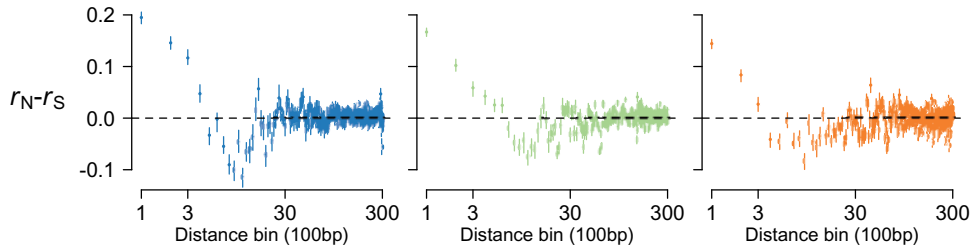

FIG. 2. NSyn couplings stronger than Syn couplings between neighboring SNPs. We binned SNP pairs by 100-bp distance intervals (bin 1 contained SNP distances between (0, 100], bin 2 contained those between (100, 200], etc.) and compared r_N with r_S for pairs within the same distance bin downsampling the category that had fewer observations (i.e., NSyn or Syn) 100 times. Dots show the median value of resampled replicates, and the vertical lines connect the 5% and 95% quantiles. Results are shown for UK (blue), NZ (green), and US (orange).

$r_N - r_S$ was significantly positive ($P \leq 5.9 \times 10^{-7}$ for all three data sets; supplementary table S1, Supplementary Material online).

A more precise approach to control for allele frequency differences would be to compare NSyn and Syn SNP pairs that have exactly matching frequencies, as opposed to binning alleles into 10% frequency intervals. This approach discards a substantial number of allele pairs and may limit the ability to detect significant differences (particularly in less diverse species such as *N. gonorrhoeae*). Nonetheless, using exactly matching sets of NSyn and Syn common alleles within 300 bp in our largest sample from UK, we again found an excess of NSyn couplings ($r_N - r_S$=0.012, with values ranging from 0.0006 to 0.025 after subsampling).

Additionally, we further controlled for recombination by categorizing genes by their density of DNA uptake sequences (DUSs; supplementary fig. S3*B*, Supplementary Material online), 12-bp motifs that dramatically increase the uptake of exogenous DNA for homologous recombination in *N. gonorrhoeae* (Goodman and Scocca 1988; Elkins et al. 1991; Ambur et al. 2007), such that genes bearing these motifs are significantly more likely to have their homologous alleles collected by the cell for recombination. Calculating $r_N - r_S$ only for genes that have low, medium, or high densities of DUSs showed similar results (supplementary fig. S3*C*, Supplementary Material online), and meta-analysis across all DUS categories again showed that $r_N - r_S$ was significantly positive ($P \leq 3.3 \times 10^{-10}$ for all three data sets; supplementary table S2, Supplementary Material online). We also found similar results when we simultaneously controlled for both allele frequency and DUS density by binning alleles into 10% frequency intervals within each DUS density category ($P \leq 2.6 \times 10^{-4}$; supplementary table S3, Supplementary Material online).

NSyn Couplings Driven by Diversity Hotspots

To further explore whether NSyn coupling linkage is shaped by positive selection, we categorized core genes by NSyn SNP density (into five quintiles), which reflects the dominant mode of selection: Genes with fewer NSyn SNPs, which predominately experience negative selection, had less 0-fold-degenerate (0D) diversity compared with 4-fold-degenerate (4D) diversity, whereas genes with many NSyn SNPs, which likely experience positive selection, had 0D/4D ratios near or above 1 (fig. 3*B*; see Materials and Methods for calculation of 0D and 4D). These genes with high NSyn SNP densities were diversity hotspots, as Syn density also increased with NSyn density (supplementary fig. S4, Supplementary Material online).

Although we observed an overall excess of NSyn couplings between neighboring alleles (figs. 2 and 3*A*), we found that the genes containing the most NSyn SNPs (top quintile) drive this overall pattern of positive $r_N - r_S$ (fig. 3*C*). The localization of NSyn couplings to high-diversity genes suggests that balancing selection may be involved, as this increases local SNP densities by maintaining haplotypes at selected and linked loci. Alternatively, although directional selection typically eliminates diversity, we may observe local increases in SNP density if it is currently spreading horizontally transferred alleles from other species.

We used *fwdpp* (Thornton 2014) to simulate these evolutionary scenarios that might create an excess of NSyn couplings within diversity hotspots, and indeed adaptive interspecies HGT and balancing selection created positive $r_N - r_S$ between neighboring alleles (fig. 3*D*). For simulations of HGT with neutral effects, $r_N - r_S$ for common alleles within 300 bp was significantly positive ($\alpha = 0.05$) for 5.1% of replicates, compared with 41% of replicates for simulations of adaptive HGT. For simulations of balancing selection, 100% of replicates produced significantly positive $r_N - r_S$ (see supplementary methods, Supplementary Material online). We also calculated $r_N - r_S$ from simulations of positive and negative directional selection in a single population, with and without positive epistasis (see supplementary results, Supplementary Material online, for more information about simulations). Although these evolutionary scenarios are not expected to increase diversity, they also did not create a significant excess of NSyn couplings for proximate alleles within the parameter space we explored (but see Discussion for limitations). Thus, accounting for interspecies HGT may provide additional insight into the mechanisms underlying NSyn coupling linkage, especially considering that *N. gonorrhoeae* is known to recombine with closely related species (Spratt et al. 1992; Feil et al. 1996; Hanage et al. 2005; Corander et al. 2012; Ezewudo et al. 2015; Wadsworth et al. 2018).

NSyn Coupling Linkage Associated with Interspecies HGT

We assembled genomic data sets for three closely related species: *N. meningitidis*, *N. polysaccharea*, and *N. lactamica*

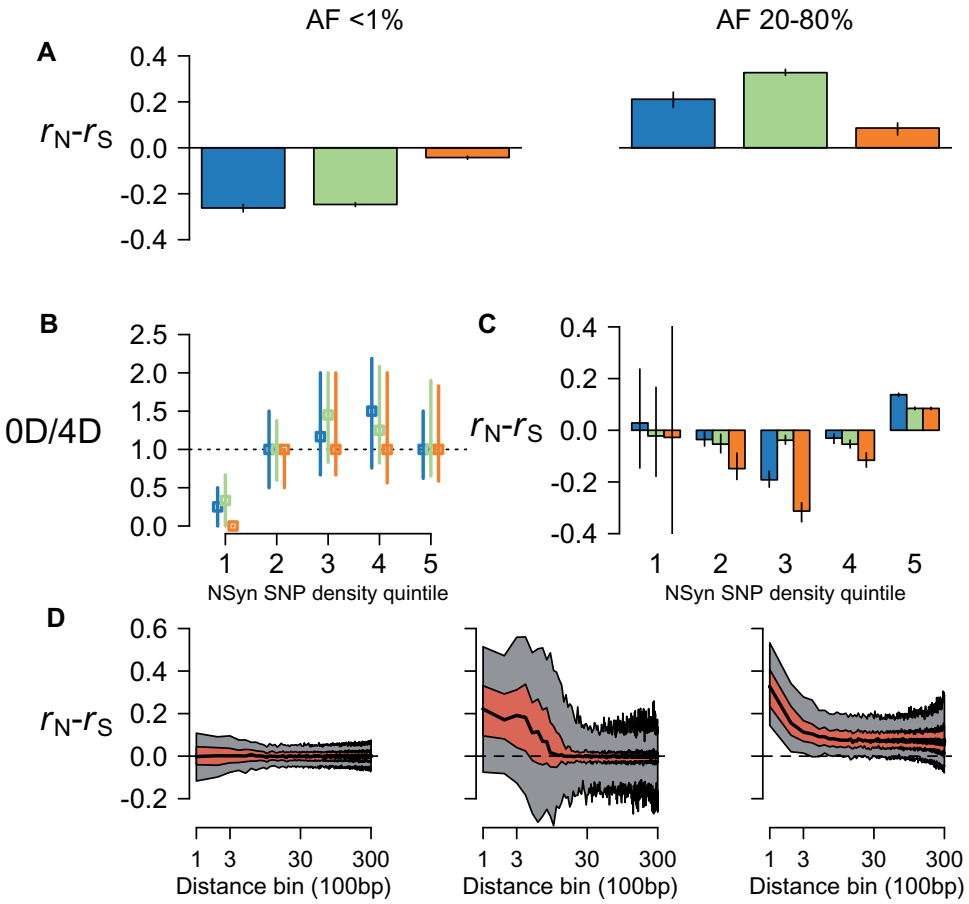

FIG. 3. Greater NSyn associations for intermediate-frequency SNPs. (*A*) Focusing only on SNPs within 300 bp, we calculated $r_N - r_S$ for SNPs with an allele frequency (AF) <1% (left) or between 20% and 80% (right). (*B*) The median ratio of diversity at 0- and 4-fold degenerate sites (OD and 4D, respectively; *y*-axis) for genes in higher NSyn SNP density quintiles was near 1, and vertical lines indicate the interquartile range. (*C*) $r_N - r_S$ was significantly >0 only for genes in the top quintile for NSyn SNP density. Results are shown for UK (blue), NZ (green), and US (orange). (*D*) We simulated interspecies recombination with beneficial mutations that have either nearly neutral ($Ns = 0.1$; left) or intermediate ($Ns = 25$; middle) effect sizes. We also simulated spatially variable selection with two demes (right) where mutations are beneficial in one deme but deleterious in the other ($Ns = \pm 4$). The central black line represents the median $r_N - r_S$, the red region spans the interquartile range of medians, and the gray region spans the 5–95% quantile of medians across 300 replicates.

(see Materials and Methods section) to characterize genome-wide allele-sharing between species. A Neighbor-Joining tree of all four species agreed with the evolutionary relationships described in previous studies (fig. 4A; Bennett et al. 2013, 2014). Figure 4B shows that intragenic diversity in *N. gonorrhoeae* was highly variable along the genome, with regions of high SNP density punctuated by regions containing few SNPs. When we compared *N. gonorrhoeae* with *N. meningitidis* (fig. 4B), peaks in SNP density visually corresponded with more shared polymorphism and less monophyletic genealogies, as measured by the genealogical sorting index (gsi), which ranges from 0 (completely mixed) to 1 (reciprocal monophyly). Indeed, when we binned genes by NSyn SNP density, those with higher densities tended to have more shared polymorphism with *N. meningitidis* and lower values of gsi (fig. 4C). We observed similar patterns when comparing *N. gonorrhoeae* with *N. polysaccharea* and *N. lactamica*, although as expected, the range of gsi values became more skewed toward 1 as genetic divergence between species increased (fig. 4B and supplementary fig. S5, Supplementary Material online).

Although interspecies HGT explains these trends, incomplete lineage sorting (ILS) may also contribute to shared ancestry, especially considering that *N. meningitidis* and *N. gonorrhoeae* are sister species. However, although HGT between diverged populations may create coupling linkage, as we observed in *N. gonorrhoeae* (fig. 4D), we show through simulations that ILS does not produce this pattern (supplementary fig. S6, Supplementary Material online). Genes with more NSyn SNPs also tended to have more DUSs (fig. 4E), which are known to significantly enhance transformation, and all *Neisseria* species analyzed here share the same DUS (Frye et al. 2013).

In summary, genes with a high density of NSyn SNPs not only had a significant excess of NSyn coupling linkage (fig. 3C) but also displayed evidence of interspecies HGT (fig. 4). An excess of NSyn couplings was also directly related to interspecies HGT, as individual genes in *N. gonorrhoeae* in which $r_N - r_S > 0.2$ also had lower values of gsi and more shared

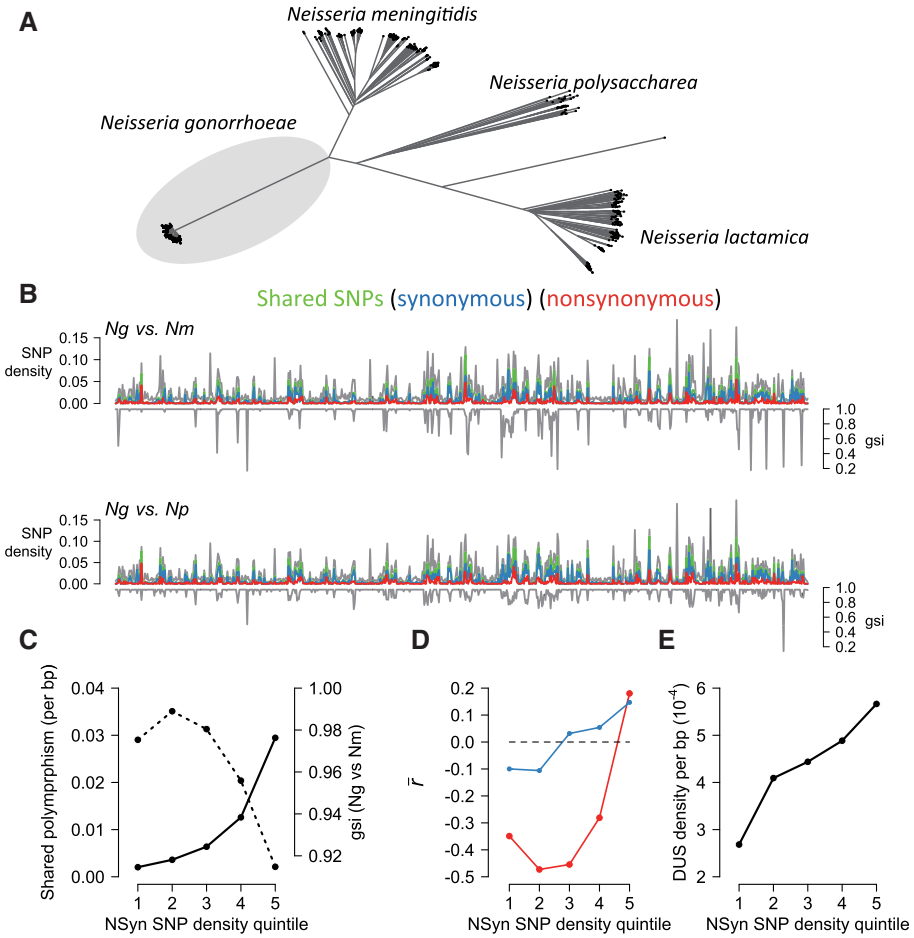

FIG. 4. Gene diversity correlated with interspecific-shared polymorphism and lower gsi values. (*A*) Unrooted Neighbor-Joining tree constructed from pairwise distances for three *Neisseria* species: *N. gonorrhoeae* (*Ng*), *N. meningitidis* (*Nm*), and *N. polysaccharea* (*Np*). (*B*) Diversity in *Ng* (only UK data set shown), quantified as SNP density (gray lines in top panels), for each gene showed that those with many SNPs also share many of these polymorphisms (green lines) with *Nm* (upper) or *Np* (lower). Blue and red lines represent shared Syn and NSyn SNPs, respectively. Genes with many SNPs also had lower values of gsi. (*C*) Categorizing genes into five NSyn SNP density quintiles showed that those with more diversity have more shared polymorphism (solid line) and lower values of gsi (dotted line). Genes with higher NSyn SNP densities also had higher mean values of $\bar{r}$ (*D*), as measured between Syn SNPs (blue line) or NSyn SNPs (red line), and more DUSs (*E*). In (*B*), genes are ordered in these plots according to their relative position in the FA1090 reference genome for *Ng*.

polymorphism with the three close relatives included here (supplementary fig. S7, Supplementary Material online), as well as slightly higher levels of intragenic DUSs (0.00059/bp vs. mean of 0.00044 for all genes). Although gsi uses gene phylogenies to measure shared ancestry, a complementary, higher resolution analysis using fastGEAR (Mostowy et al. 2017) also showed how ancestry changes across the length of genes with high $r_N - r_S$ values, revealing tracts of DNA from other species or entire alleles that have no identifiable DNA from *N. gonorrhoeae* (supplementary fig. S8, Supplementary Material online). Collectively, these observations highlight the potential role of interspecies HGT in driving NSyn couplings within highly diverse genes under positive selection.

Outlier Genes with an Excess of NSyn Coupling Linkage

The distribution of $r_N - r_S$ calculated for individual genes was skewed toward positive values, with most genes having near zero values (fig. 5). Although $r_N - r_S$ outlier genes generally had more NSyn polymorphism, of the 29 outlier genes we detected in at least one data set ($r_N - r_S > 0.2$, Materials and Methods), only one exhibited $dN/dS > 1$ (supplementary table S4, Supplementary Material online). Most of these outlier genes are distant from one another (>10 kb) according to their position within the FA1090 reference genome, but three pairs of genes are <3 kb apart and may be linked (supplementary table S4, Supplementary Material online).

Twenty of these genes had annotation information and over half (14) were involved in metabolic processes (supplementary table S4, Supplementary Material online), suggesting that selection on metabolism creates an important component of haplotype structure in *Neisseria*. Of the other outliers, two were membrane proteins, and another was *mtrE*, a candidate for adaptive HGT, as it is part of an operon encoding an efflux pump that has acquired antibiotic resistance via recombination with several closely related species (Wadsworth et al. 2018). Many of these $r_N - r_S$ outlier genes also had a large excess of intermediate-frequency variants

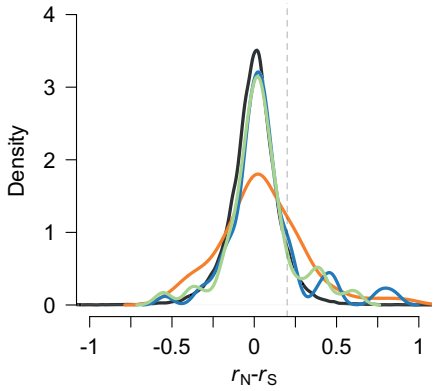

FIG. 5. Distribution of intragenic $r_N - r_S$ values. Values of $r_N - r_S$ by gene showed a positive-skewed distribution with an excess of NSyn couplings compared with neutral simulations of a single population (black). The dashed gray line indicates the threshold used for $r_N - r_S$ outliers. For each gene with at least five Syn and five NSyn SNPs, $r_N - r_S$ was calculated for derived alleles with frequencies above 5%, Results are shown for UK (blue), NZ (green), and US (orange).

(supplementary table S4, Supplementary Material online), a pattern expected under balancing selection. Of these, one encodes a membrane protein, a category that is frequently documented to experience such selective pressures (Gupta et al. 1996), and two others encode metabolic proteins that act in different parts of the same amino sugar and nucleotide sugar pathway. These balancing selection candidates also did not exhibit evidence of interspecies HGT according to gsi values (at least for the three outgroup species used here), such that an excess of NSyn couplings may be driven solely by balancing selection on intraspecific diversity.

Discussion

Neisseria gonorrhoeae is a highly recombining human pathogen that has stably colonized multiple ecological niches, including the urogenital tract, the oropharynx, and the rectum, and has evolved resistance to all clinically important antibiotics used to treat infection. Given the pervasiveness of recombination within and between *Neisserial* species, analyses that jointly study selection and recombination could reveal novel insights into the biology of the *Neisseria* that other approaches will not detect. Using patterns of directional linkage between derived mutations in *N. gonorrhoeae*, we show that associations between neighboring polymorphisms are shaped by a form of selection that creates stronger coupling linkage between NSyn alleles compared with Syn alleles ($r_N - r_S > 0$). Other summary statistics show at least some of this positive selection may act on alleles horizontally transferred from closely related species, and simulations with adaptive HGT between diverged populations or balancing selection produce similar patterns.

These two forms of selection, which are not mutually exclusive, inflate NSyn couplings between close alleles in different ways. In the case of balancing selection, which maintains allelic variation in a large panmictic population that encounters heterogeneous environments, NSyn couplings represent allelic combinations that are beneficial in one environment but neutral or deleterious in others. Although recombination occurs between all lineages and breaks linkage, selection maintains preferred allele combinations if it is sufficiently stronger than recombination, a scenario that becomes increasingly likely for neighboring alleles that experience less recombination (Wiuf and Hein 2000; Lin and Kussell 2019). In the case of adaptive HGT, linked beneficial NSyn alleles are introduced on short recombination tracts that can rapidly increase in frequency, which makes them younger (providing less opportunity for recombination to break their linkage) than Syn alleles that rose to the same frequency via drift. This process of adaptive HGT with linked NSyn mutations contrasts starkly with beneficial NSyn mutations that randomly arise within a species, as these typically display repulsion linkage from clonal interference due to limited recombination (supplementary fig. S18, Supplementary Material online; Hill and Robertson 1968). Importantly, simulations of interspecies HGT involving neutral mutations do not tend to produce positive values of $r_N - r_S$ (fig. 3D).

Although our simulations of negative selection with positive (antagonistic) epistasis did not produce an excess of NSyn couplings within genes, results may differ with very strong epistasis in which NSyn alleles are highly deleterious but perfectly compensate one another (Callahan et al. 2011). However, in *N. gonorrhoeae* the excess of NSyn couplings is localized within a small fraction of highly diverse genes under weaker selective constraints (fig. 3B), and many of these genes display an excess of intermediate-frequency alleles and have biological functions compatible with balancing selection or adaptive HGT. Negative selection with epistasis—which does not produce diversity hotspots or an enrichment of intermediate-frequency alleles—within a minority of genes is thus an unlikely explanation.

The $r_N - r_S$ metric developed here serves as a complementary way to study selection in bacterial and other sexual organisms, as it detects signatures of selection that differ from those identified by other methods, such as d*N*/d*S* or metrics that compare diversity with divergence (McDonald and Kreitman 1991; Stoletzki and Eyre-Walker 2011). Intriguingly, an excess of couplings between close Syn alleles within ~300 bp, as in figure 1B, has also been documented in a highly recombining, thermophilic *Synechococcus* species (Rosen et al. 2015), and patterns of linkage between NSyn alleles suggest that positive selection may maintain allelic combinations between diverged groups (or "clouds"; Rosen et al. 2018). These findings also have important implications for estimating recombination parameters from bacterial genomic data using neutral models (Lin and Kussell 2019), as selection may shape short-range patterns of linkage.

Although numerous unknown selective pressures may be driving the maintenance of diversity in *N. gonorrhoeae*, epidemiological data provide promising clues. *Neisseria gonorrhoeae* thrives within different sexual networks in which lineages are transmitted through diverse habitats that may present distinct challenges. For instance, in networks involving men who have sex with men, *N. gonorrhoeae* has been isolated from the urethra, rectum, and oropharynx, and for networks involving women, its niches include the vagina/

cervix (Grad et al. 2014; De Silva et al. 2016; Lee et al. 2018; Sánchez-Busó et al. 2018). Adaptive HGT and the coexistence of diversity could thus reflect spatially or temporally variable selection from lineages sojourning in these distinct niches.

Horizontally transferred diversity could also represent recently acquired alleles that are unconditionally beneficial and currently spreading through the entire species. *Neisseria gonorrhoeae* has substantially less diversity than all of its close relatives (fig. 4A), which may reflect a genetic bottleneck that accompanied recent speciation, such as a single *N. meningitidis* lineage giving rise to the present-day species. Severe bottlenecks may fix rare, deleterious variation present in the ancestral population, such that adaptive HGT may simply reflect the acquisition of beneficial wild-type alleles (Kim et al. 2018).

Additional analyses offer the opportunity for a more complete understanding of genome-wide evolution in *N. gonorrhoeae*. Here, we only consider core genes and focus on contemporaneous samples by subsampling larger data sets to obtain isolates from a restricted time period (e.g., a single year). Analyses that include longitudinal samples will help understand the temporal dynamics of mutations and may be particularly useful for establishing types of positive selection, for instance, whether directional selection on horizontally transferred haplotypes or balancing selection shapes patterns of variation at a specific locus, as the latter scenario will maintain polymorphisms over time. Moreover, a deeper understanding of the other commensal species within *Neisseria*, such as any species *N. gonorrhoeae* may recombine with, will shed additional light on *N. gonorrhoeae* biology and speciation processes within the genus.

Materials and Methods

Sequence Data Processing

We reanalyzed three previously published sequencing data sets of *N. gonorrhoeae* isolates collected from Brighton (UK; De Silva et al. 2016), NZ (Lee et al. 2018), and US (Grad et al. 2014, 2016). As sequencing errors may give rise to structure in genomic data sets (and thus positive $\bar{r}$) if subsets of isolates are more likely to have erroneous nucleotides, we took many precautions to ensure high-quality DNA alignments for downstream analysis.

We analyzed the quality of raw reads using FastQC (Andrews 2010), and samples with GC content that differed more than 2.5 SD from the mean were not included in analyses (three isolates in the data set from Brighton). These reads were used to create de novo assemblies using SPAdes (v. 3.11; Bankevich et al. 2012), and contigs were joined into larger scaffolds using SSPACE (Boetzer et al. 2011). We mapped raw reads back to these assemblies using SMALT (v. 0.7.6), using only those reads in which at least 95% of bases successfully mapped. With these mapped alignments, we used Pilon (v. 1.13; Walker et al. 2014) to correct any nucleotides not supported by the raw read data. For a maximum of four iterations, we repeated this process of mapping reads back to the de novo assembly and correcting nucleotides, unless Pilon made no changes to the updated assembly.

Using these polished assemblies along with information from the final mapped alignment, we marked nucleotide positions as "N" if they met any of the following criteria: 1) alignment depth <5 bases, 2) mean base quality <20, 3) mean mapping quality <20, 4) alignment depth more than 1.8 times the median depth across all positions, 5) positions flagged as "Amb" by Pilon that have significant evidence of more than one allele, or 6) positions in which a second allele was present with a frequency $>10\%$ of all base calls. These positions marked as "N" were excluded from downstream analyses. In addition, we discarded all contigs that were <300 bp or had more than 50% of their sites masked by the filtering step above (primarily plasmids with very high sequencing depth).

Gene Annotation and Alignment

We annotated assemblies with Prokka (Seemann 2014) using the proteome of the FA1090 reference genome. We then used Roary (Page et al. 2015) to identify core genes present in all *N. gonorrhoeae* isolates, defining orthologous genes as having at least 90% amino acid similarity. This threshold tends to misclassify core genes with very diverged alleles as multiple accessory genes (Ding et al. 2018), such that we may miss those that have admixed with distant species. However, we preferred a conservative set of core genes for accurate estimates of linkage, as erroneous clustering of alleles may give rise to artifactual associations.

We then realigned these core genes with PAGAN (Löytynoja et al. 2012), a phylogeny-aware multiple sequence aligner. PAGAN performs an amino acid alignment that helps maintain the reading frame of diverged sequences, which is required for accurately labeling polymorphisms as Syn or NSyn. All position information between polymorphic sites was derived from the relative positions of genes in the FA1090 reference genome used to annotate de novo assemblies, not from a reference-based DNA alignment. Although rearrangements may occur within *N. gonorrhoeae*, synteny between genomes likely extends beyond 3 kb (supplementary fig. S9, Supplementary Material online), the distance around which r^2 and r approach near-zero values. Moreover, codons in the FA1090 reference sequence were used to ascertain the functional effect of polymorphisms, that is, whether they were Syn or NSyn. For all downstream analyses, we excluded COGS that were present in multiple copies in at least one individual and all alignments that contained multiple premature stop codon polymorphisms.

Polarizing Mutations

For analyses that required polarized mutations (e.g., calculating r), we used progressiveMauve (Darling et al. 2010) to align the FA1090 reference genome to an *N. meningitidis* outgroup sequence to infer the derived and ancestral state of each biallelic polymorphism within *N. gonorrhoeae*. All analyses were done using the α14 *N. meningitidis* reference as the outgroup sequence, as it is the closest known reference to *N. gonorrhoeae* (Budroni et al. 2011). However, as this

reference sequence may have acquired derived mutations since its divergence from *N. gonorrhoeae*, we also polarized mutations using one *N. meningitidis* (α14) and one *N. polysaccharea* (Short Read Archive accession number ERR976854) sequence. We only considered positions in which both outgroup sequences had the same nucleotide and *N. gonorrhoeae* had a biallelic polymorphism, with one of the alleles also found in the outgroup. The outgroup consensus allele served as the ancestral state. This method of polarization gave highly similar results (supplementary fig. S10, Supplementary Material online). We also note that we included the FA1090 reference sequence when clustering COGs with Roary and realigning with PAGAN (above), so that we could map positions within gene alignments to those in the reference used in the multispecies Mauve alignment used to polarize mutations. However, this reference sequence was excluded from analyses of diversity and linkage within each of the three data sets.

Data Filtration

With these core genome alignments, we calculated the number of pairwise SNP differences between alignments using Disty McMatrixface (https://github.com/c2-d2/disty) and down sampled closely related isolates in order to exclude those from the same transmission chain. Clusters of closely related genomes could arise from reporting bias within contact networks, as isolates were sampled from symptomatic individuals that visited sexual health clinics. Specifically, we down sampled clusters of isolates that were identical (0 SNP differences) or separated by <6 SNPs to a randomly selected isolate within that cluster. We used the 6-SNP threshold because De Silva et al. (2016) showed that isolates collected from known contact networks in low-transmission settings had fewer than six SNP differences. For either SNP threshold, patterns of pairwise $\bar{r}$ looked highly similar (supplementary fig. S11, Supplementary Material online), and all analyses shown here were performed on alignments down sampled according to the 6-SNP threshold. We also excluded isolates that were collected from the same patient over time (UK data set), although these samples would have also been excluded using the SNP thresholds above.

Lastly, to avoid any potential temporal structure that may inflate $\bar{r}$, we analyzed only isolates collected in 2013 from the UK data, and only those collected in 2010–2011 from the US data set. All isolates in the NZ data set collected from 2014 to 2015 were analyzed. Overall, we analyzed 214 sequences from UK, 149 from US, and 148 from NZ.

Linkage Statistics

We primarily measured linkage using r, where $r = D/\sqrt{p_i(1-p_i)p_j(1-p_j)}$ and $D = p_{ij} - p_i p_j$, or the probability that derived mutations at site i and j occur together minus the random expectation based on their individual frequencies (Hill and Robertson 1968). Although r and r^2 are commonly used to quantify linkage, their maximum values depend on the degree of symmetry between the allele frequencies of a pair of loci under consideration (VanLiere and Rosenberg 2008). However, the range of values of another linkage statistic, D' (Lewontin 1964), is independent of allele frequencies (Hedrick 1987). We repeated our $r_N - r_S$ analyses (figs. 2 and 3C) using D' and found qualitatively similar results: Proximate derived NSyn alleles had more coupling linkage than derived Syn alleles separated by similar distances, and this observation was driven by genes with the highest density of NSyn SNPs (supplementary fig. S12, Supplementary Material online).

We also note that for analyses in figure 4D, although theory has shown that $\bar{r} = 0$ for a panmictic species (Hill and Robertson 1968; Ohta and Kimura 1969, 1971), we confirmed with simulations that conditioning on SNP density does not alter this expectation (supplementary fig. S13, Supplementary Material online), such that genes with many SNPs are not expected to have higher $\bar{r}$, as we observe.

All data for *N. gonorrhoeae*, including gene alignments and input files from other software, and documented Perl scripts used in these analyses are freely available on github (https://github.com/brian-arnold/NgonorrhoeaeLinkageGenomics).

Diversity Statistics

We determined 0D and 4D sites using the *N. gonorrhoeae* FA1090 reference genome along with the genetic code. A site was recorded as 0D if every possible nucleotide substitution changed the amino acid present in the reference sequence and as 4D if every nucleotide substitution did not change the amino acid. 0D (4D) diversity was then calculated as the fraction of 0D (4D) sites within a gene that contained an SNP.

Significance Tests

When accounting for allele frequencies or DUS densities (supplementary fig. S2 and tables S1–S3, Supplementary Material online), we tested whether $r_N - r_S$ was significantly positive within each bin or category using the Wilcoxon rank-sum test ("Stats" package in R). To meta-analyze P values across bins, we used Stouffers Z-score method ("metap" package in R; Liptak 1958). We also applied these tests to simulated data (supplementary methods, Supplementary Material online).

Interspecies Analysis

To study shared ancestry between *N. gonorrhoeae* and its close relatives, we downloaded all assemblies from NCBI or raw read data from the Short Read Archive for *N. meningitidis*, *N. polysaccharea*, and *N. lactamica* on October 19, 2017. We created de novo assemblies from these raw reads using SPAdes as above. We then only used assemblies that had N50 >10 kb, no more than 150 contigs, and at least one contig that was at least 300 kb. We also excluded several assemblies that were very diverged from the majority of samples for that species according to a visualization of Neighbor-Joining trees (1 for *N. meningitidis*, 15 for *N. lactamica*, and 4 for *N. polysaccharea*); these highly diverged isolates may have bad assemblies or were incorrectly labeled. In total we used 431 *N. meningitidis*, 326 *N. lactamica*, and 37 *N. polysaccharea* assemblies, and the accession numbers for these may be found in supplementary table S5, Supplementary Material online.

We then mapped sequences from each outgroup species to the *N. gonorrhoeae* FA1090 reference genome. However,

instead of mapping raw reads, we first used them to create de novo assemblies (using SPAdes, as above) for each isolate and then mapped scaffolds from these assemblies to the reference using progressiveMauve (Darling et al. 2010). We opted for this approach because long scaffolds have more information about mapping position than short sequencing reads, and microsynteny among the four species extends beyond the length of a gene (supplementary fig. S9, Supplementary Material online). For each gene we previously identified as "core" within the *N. gonorrhoeae* data sets, we extracted the sequences from these progressiveMauve alignments and incorporated them into existing core gene alignments with MAFFT (using the –add and –keeplength options; Katoh and Frith 2012). We excluded genes that were present in fewer than 20% of isolates in the outgroup species or with alignments in which over 50% of positions were gaps. Although these alignments were directly used to quantify shared ancestry in terms of shared SNP density, we also constructed a multispecies phylogeny for each gene using RAxML (v. 8.1.5; Stamatakis 2014) with 20 bootstrap replicates under the GTRCAT model of rate heterogeneity. These phylogenies were used to calculate gsi (Cummings et al. 2008) with the genealogicalSorting R package, and we took the mean gsi value across all 20 bootstrap replicates for leaves labeled as *N. gonorrhoeae*.

Forward-Time Simulations

To simulate bacterial evolution, we used *fwdpp* (Thornton 2014), a library of C++ functions that abstracts the essential tasks of forward-time simulations such that one may use them to design custom simulators. We modified the code to accommodate haploid populations and also made custom functions to simulate homologous recombination (i.e., gene conversion) and various types of selection. With *fwdpp*, we simulate Wright–Fisher metapopulations under an infinite-sites mutation model. For more information about details of simulation parameters and results, and the calculation of individual fitness, please see supplementary methods, Supplementary Material online. Source code is available at https://github.com/brian-arnold/NgonorrhoeaeLinkageGenomics.

DUS Identification and Density Estimation

To locate DUSs within *Neisseria* genomes, we searched sequences for the degenerate 10-bp motif 5′-GCCGTCTGAA-3′, allowing one nucleotide to vary within the first three or last two positions because mutations within positions 4–8 may result in highly reduced rates of uptake (Frye et al. 2013). We searched both forward and reverse strands. When counting DUS densities within genes or their flanking regions (200 bp), we collapsed DUSs within 300 bp into a single observation: Although increasing the number of DUSs generally increases DNA uptake and transformation, those separated by short distances exhibit interference (Ambur et al. 2012). DUS density per gene (including flanks) is highly variable (supplementary fig. S2*B*, Supplementary Material online). When controlling for relative DUS density while calculating $r_N - r_S$, we created three categories (high, medium, and low) but excluded genes with either no DUSs or very high DUS densities (top 15%) to avoid categories with highly variable densities (supplementary fig. S2*B*, Supplementary Material online).

dN/dS Analysis

Using *omegaMap* (Wilson and McVean 2006), we calculated dN/dS for genes that also had high values of $r_N - r_S$ (supplementary table S4, Supplementary Material online). For information about the parameters used in this analysis, please see the supplementary methods, Supplementary Material online.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank Kevin Thornton for his generous help with *fwdpp*. B.J.A. was supported by a postdoctoral fellowship1 F32 GM120839-01, C.B.W. and Y.H.G. were supported by the Richard and Susan Smith Family Foundation and National Institutes of Health (Grant No. R01 AI132606). J.C. was funded by the ERC (Grant No. 742158), and W.P.H. was funded by National Institutes of Health (Grant No. U54 GM088558). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The computations in this study were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. The authors declare no competing interests.

References

Ambur OH, Frye SA, Nilsen M, Hovland E, Tønjum T. 2012. Restriction and sequence alterations affect DNA uptake sequence-dependent transformation in *Neisseria meningitidis*. *PLoS One* 7(7):e39742–12.

Ambur OH, Frye SA, Tønjum T. 2007. New functional identity for the DNA uptake sequence in transformation and its presence in transcriptional terminators. *J Bacteriol*. 189(5):2077–2085.

Arnold BJ, Gutmann MU, Grad YH, Sheppard SK, Corander J, Lipsitch M, Hanage WP. 2018. Weak epistasis may drive adaptation in recombining bacteria. *Genetics* 208(3):1247–1260.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 19(5):455–477.

Bennett JS, Jolley KA, Maiden MCJ. 2013. Genome sequence analyses show that *Neisseria oralis* is the same species as "*Neisseria mucosa* var. *h*eidelbergensis." *Int J Syst Evol Microbiol*. 63(Pt 10):3920–3926.

Bennett JS, Watkins ER, Jolley KA, Harrison OB, Maiden MCJ. 2014. Identifying *Neisseria* species by use of the 50S Ribosomal protein L6 (rplF) gene. *J Clin Microbiol*. 52(5):1375–1381.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578–579.

Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV. 2011. *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci U S A*. 108(11):4494–4499.

Callahan B, Neher RA, Bachtrog D, Andolfatto P, Shraiman BI. 2011. Correlated evolution of nearby residues in Drosophilid proteins. PLoS Genet. 7(2):e1001315.

Charlesworth B, Nordborg M, Charlesworth D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res*. 70(2):155–174.

Corander J, Connor TR, O'Dwyer CA, Kroll JS, Hanage WP. 2012. Population structure in the *Neisseria*, and the biological significance of fuzzy species. *J R Soc Interface*. 9(71):1208–1215.

Cui Y, Yang X, Didelot X, Guo C, Li D, Yan Y, Zhang Y, Yuan Y, Yang H, Wang J, et al. 2015. Epidemic clones, oceanic gene pools, and eco-LD in the free living marine pathogen *Vibrio parahaemolyticus*. *Mol Biol Evol*. 32(6):1396–1410.

Cummings MP, Neel MC, Shaw KL. 2008. A genealogical approach to quantifying lineage divergence. *Evolution (NY)* 62:2411–2422.

Darling AE, Mau B, Perna NT. 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* 5(6):e11147.

De Silva D, Peters J, Cole K, Cole MJ, Cresswell F, Dean G, Dave J, Thomas DR, Foster K, Waldram A, et al. 2016. Whole-genome sequencing to determine transmission of *Neisseria gonorrhoeae*: an observational study. *Lancet Infect Dis*. 16(11):1295–1303.

Ding W, Baumdicker F, Neher RA. 2018. panX: pan-genome analysis and exploration. *Nucleic Acids Res*. 46:1–12.

Elkins C, Thomas CE, Seifert HS, Sparling PF. 1991. Species-specific uptake of DNA by gonococci is mediated by a 10-base-pair sequence. *J Bacteriol*. 173(12):3911–3913.

Eshel I, Feldman MW. 1970. On the evolutionary effect of recombination. *Theor Popul Biol*. 1(1):88–100.

Ezewudo MN, Joseph SJ, Castillo-Ramirez S, Dean D, del Rio C, Didelot X, Dillon J-A, Selden RF, Shafer WM, Turingan RS, et al. 2015. Population structure of *Neisseria gonorrhoeae* based on whole genome data and its relationship with antibiotic resistance. *PeerJ*. 3:e806.

Feil E, Zhou J, Smith JM, Spratt BG. 1996. A comparison of the nucleotide sequences of the adk and recA genes of pathogenic and commensal *Neisseria* species: evidence for extensive interspecies recombination within adk. *J Mol Evol*. 43(6):631–640.

Felsenstein J. 1974. The evolutionary advantage of recombination. II. Individual selection for recombination. *Genetics* 83:845–859.

Frye SA, Nilsen M, Tønjum T, Ambur OH. 2013. Dialects of the DNA uptake sequence in Neisseriaceae. *PLoS Genet*. 9(4):e1003458.

Goodman SD, Scocca JJ. 1988. Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc Natl Acad Sci U S A*. 85(18):6982–6986.

Grad YH, Harris SR, Kirkcaldy RD, Green AG, Marks DS, Bentley SD, Trees D, Lipsitch M. 2016. Genomic epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and fluoroquinolones in the United States, 2000-2013. *J Infect Dis*. 214(10):1579–1587.

Grad YH, Kirkcaldy RD, Trees D, Dordel J, Harris SR, Goldstein E, Weinstock H, Parkhill J, Hanage WP, Bentley S, et al. 2014. Genomic epidemiology of *Neisseria gonorrhoeae* with reduced susceptibility to cefixime in the USA: a retrospective observational study. *Lancet Infect Dis*. 14(3):220–226.

Gupta S, Maiden MCJ, Feavers IM, Nee S, May RM, Anderson RM. 1996. The maintenance of strain structure in populations of recombining infectious agents. *Nat Med*. 2(4):437–442.

Hanage WP, Fraser C, Spratt BG. 2005. Fuzzy species among recombinogenic bacteria. *BMC Biol*. 3(1):6–7.

Hedrick PW. 1987. Gametic disequilibrium measures: proceed with caution. *Genetics* 117(2):331–341.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet*. 38(6):226–231.

Hughes AL. 2005. Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* 169(2):533–538.

Katoh K, Frith MC. 2012. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* 28(23):3144–3146.

Kim BY, Huber CD, Lohmueller KE. 2018. Deleterious variation shapes the genomic landscape of introgression. *PLoS Genet*. 14(10):e1007741.

Lee RS, Seemann T, Heffernan H, Kwong JC, Gonçalves da Silva A, Carter GP, Woodhouse R, Dyet KH, Bulach DM, Stinear TP, et al. 2018. Genomic epidemiology and antimicrobial resistance of *Neisseria gonorrhoeae* in New Zealand. *J Antimicrob Chemother*. 73(2):353–364.

Lewontin RC. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49(1):49–67.

Lin M, Kussell E. 2019. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat Methods*. 16(2):199–204.

Liptak T. 1958. On the combination of independent tests. *Magy Tud Akad Mat Kut Int Kozl*. 3:171–197.

Löytynoja A, Vilella AJ, Goldman N. 2012. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics* 28(13):1684–1691.

Martin G, Otto SP, Lenormand T. 2006. Selection for recombination in structured populations. *Genetics* 172(1):593–609.

McDonald J, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351(6328):652–654.

McVean GAT, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* 155(2):929–944.

Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. 2017. Efficient inference of recent and ancestral recombination within bacterial populations. *Mol Biol Evol*. 34(5):1167–1182.

O'Rourke M, Stevens E. 1993. Genetic structure of *Neisseria gonorrhoeae* populations: a non-clonal pathogen. *J Gen Microbiol*. 139:2603–2611.

Ohta T, Kimura M. 1969. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* 63(1):229–238.

Ohta T, Kimura M. 1971. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68(4):571–580.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31(22):3691–3693.

Rosen M, Davison M, Bhaya D, Fisher DS. 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science (80-.)* 348(6238):1019–1024.

Rosen MJ, Davison M, Fisher DS, Bhaya D. 2018. Probing the ecological and evolutionary history of a thermophilic cyanobacterial population via statistical properties of its microdiversity.

Sánchez-Busó L, Golparian D, Corander J, Grad YH, Ohnishi M, et al. 2018. Antimicrobial exposure in sexual networks drives divergent evolution in modern gonococci. bioRxiv: 334847.

Schubert B, Maddamsetti R, Nyman J, Farhat MR, Marks DS. 2019. Genome-wide discovery of epistatic loci affecting antibiotic resistance in *Neisseria gonorrhoeae* using evolutionary couplings. *Nat Microbiol*. 4(2):328–338.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069.

Shapiro BJ, David LA, Friedman J, Alm EJ. 2009. Looking for Darwin's footprints in the microbial world. *Trends Microbiol*. 17(5):196–204.

Skwark M, Croucher N, Puranen S, Chewapreecha C, Pesonen M, Xu YY, Turner P, Harris SR, Beres SB, Musser JM, et al. 2016. Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genet*. 13(2):e1006508.

Smith JM, Smith NH, O'Rourke M, Spratt BG. 1993. How clonal are bacteria? *Proc Natl Acad Sci U S A*. 90(10):4384–4388.

Sohail M, Vakhrusheva OA, Sul JH, Pulit SL, Francioli LC, van den Berg LH, Veldink JH, de Bakker PIW, Bazykin GA, Kondrashov AS, et al. 2017. Negative selection in humans and fruit flies involves synergistic epistasis. *Science (80-.)* 356(6337):539–542.

Spratt BG, Bowler LD, Zhang QY, Zhou J, Smith JM. 1992. Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. *J Mol Evol*. 34(2):115–125.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Mol Biol Evol*. 28(1):63–70.
Takahasi KR, Innan H. 2008. The direction of linkage disequilibrium: a new measure based on the ancestral-derived status of segregating alleles. *Genetics* 179(3):1705–1712.
Thomson G. 1977. The effect of a selected locus on linked neutral loci. *Genetics* 85(4):753–788.
Thornton KR. 2014. A C++ template library for efficient forward-time population genetic simulation of large populations. *Genetics* 198(1):157–121.
VanLiere JM, Rosenberg NA. 2008. Mathematical properties of the r2 measure of linkage disequilibrium. *Theor Popul Biol*. 74(1):130–137.
Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. *ISME J*. 3(2):199–208.
Wadsworth CB, Arnold BJ, Sater MRA, Grad Y. 2018. Azithromycin resistance through interspecific acquisition of an epistasis-dependent efflux pump component and transcriptional regulator in *Neisseria gonorrhoeae*. *MBio*. 9(4):1–17.
Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9(11):e112963.
Wilson DJ, McVean G. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* 172(3):1411–1425.
Wiuf C, Hein J. 2000. The coalescent with gene conversion. *Genetics* 155(1):451–462.